DOCUMENT RESUME

ED 349 342 TM 019 067

AUTHOR Herman, Joan
TITLE What's Happening with Educational Assessment?
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; Southeastern Regional Vision for Education (SERVE), Tallahassee, FL.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Jun 92
CONTRACT RP91002010
NOTE 15p.; Synopsis of two presentations delivered at the Annual Meeting of the SouthEastern Regional Vision for Education (Tampa, FL, February 1992).
PUB TYPE Information Analyses (070) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Tests; *Educational Assessment; Educational Change; Educational Improvement; Educational Policy; Elementary Secondary Education; Multiple Choice Tests; Psychometrics; Research and Development; Standardized Tests; *Student Evaluation; Test Results; *Test Use; *Thinking Skills
IDENTIFIERS *Alternatives to Standardized Testing; *Reform Efforts

ABSTRACT

Using tests to reform education is not a new concept. National and state education policies reflect confidence in the power of assessment to encourage school improvement. New understandings of the nature and context of student learning are shifting the movement away from traditional multiple-choice tests to alternative assessments. The emphasis is now on assessing higher-level thinking skills rather than lower-level skills and rote memory. Assessment research indicates that educational processes can be distorted when there is a lot of pressure on test results and high-stakes consequences follow the results. Narrowing of the curriculum and adjustment of the system to protect children are two undesirable consequences. Reviewing what is happening in assessment nationally reveals the following as key players: (1) America 2000; (2) the National Education Goals Panel; (3) the National Council on Evaluation, Standards, and Student Testing; (4) the Secretary's Commission on Achieving Necessary Skills; (5) the National Standards Project; (6) the National Assessment of Educational Progress; (7) the National Assessment Governing Board; (8) the National Council of Chief State School Officers; (9) the Student Assessment Consortium; and (10) test publishers. Research and development challenges in rethinking assessment include a lack of empirical research, psychometric issues, and issues of generalizability and validity. The real consequences of reform must be examined, with regard to equity issues and the problems of calibrating a national system of tests. (SLD)

ED349342

WHAT'S HAPPENING WITH EDUCATIONAL ASSESSMENT?

June 1992

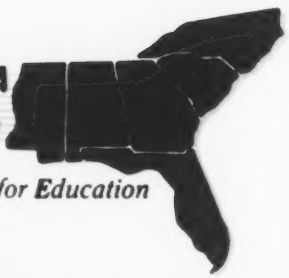
U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM019067



BEST COPY AVAILABLE

WHAT'S HAPPENING WITH EDUCATIONAL ASSESSMENT?

June 1992

Joan Herman, Ph.D.

Co-published by

Center for the Study of Evaluation
National Center for Research on Evaluation, Standards, & Student Testing
(CSE/CRESST)
University of California at Los Angeles

and

SouthEastern Regional Vision for Education
(SERVE)

Dr. Joan Herman, Associate Director of the UCLA Center for the Study of Evaluation, is one of the nation's leaders in the field of educational assessment. The following report is a synopsis of two presentations Dr. Herman delivered before the Governing Board of SERVE at its February 1992 meeting in Tampa, Florida.

Edited by Jane Menton, Research Associate for Policy, SERVE

Special thanks to Mark Heidorn, Program Director for Assessment, Florida Department of Education, and SERVE staff Dianne Wilkes, Joe Follman, and Stephanie Kadel, for their review and comments.

For additional information contact:

Center for the Study of Evaluation
UCLA Graduate School of Education
145 Moore Hall
405 Hilgard
Los Angeles, CA 90024-1522

(310) 206-1532

SERVE
345 South Magnolia Drive
Suite D-23
Tallahassee, FL 32301-2950

(800) 352-6001

This document was produced with funds from the Office of Educational Research and Improvement, U.S. Department of Education, under contract number RP91002010.

Educational assessment is currently in the process of invention--
Old models are being seriously questioned.
New models are in the process of development.
Everyone is talking about it.
Everyone is advocating it.
Lots of people are trying it.
No one really knows how to do it.
We are truly inventing the future as we go along.

Dr. Joan Herman, 1992

Why Assess?

Using tests to reform education is not a new concept. In the 1970s, it was the minimum competency movement. In the 1980's, the country experienced a rash of new and expanded test requirements as a result of the publication of *A Nation at Risk*. Today, national and state education policy reflects a continued confidence in the power of assessment to encourage school improvement, premised on the belief that:

- Good assessment sets meaningful standards to which school systems, schools, teachers, and students can aspire.
- These standards provide direction for instructional efforts and models of good practice.
- Assessment results provide feedback on instructional strengths and weaknesses and prescriptions for action at all levels of the educational system.
- Coupled with effective incentives and/or sanctions, assessment can motivate students to learn better, teachers to teach better, and schools to be more educationally effective.

At the micro-level, we assess a child because we want to know his or her capability in some larger domain. At the macro-level, we use assessment as a policy tool that serves as a multi-level intervention strategy to ensure school performance and accountability. For example: if you are a national policymaker, you can think of assessment as a way to affect and motivate all the levels below you; if you are a state policymaker, you can think of assessment as a tool for managing districts or intermediate units, schools, and teachers; if you are a school principal, you can think of assessment as a tool for managing your instructional programs, teachers, and students; if you are a teacher, you can think about assessment as a way to motivate your students and bring parents into the action.

Based on the above, assessment has the potential to be a powerful and beneficial engine of change--but in the 1990's, the train looks dramatically different. New understandings of the nature and context of student learning have shifted the movement away from traditional, multiple-choice tests to alternative assessments.

What Do We Mean by Alternative Assessment?

Assessment continues to be one of the cornerstones of educational reform in the 1990's, but today's testing proponents are taking a dramatically different tact--one that even demands new terminology. "Performance-based," "outcome-based," "hands-on," "authentic," and "alternative" assessment are some of the many terms used by educators and others who are working to affect and measure meaningful educational improvement. The types of assessment strategies now advocated take on a variety of shapes including essays, open-ended questions, demonstrations, hands-on tasks, and portfolios of work that may be accumulated over a year or more. Essentially, rather than measuring lower-level skills and rote memory (believed to be the emphasis of traditional multiple-choice testing), the emphasis now is on assessing higher-level thinking skills. While the terms may be diverse, several common threads link these alternative assessments.

Common Elements of Alternative Assessment Strategies

- ❑ Students perform, create, produce, or do something.
- ❑ Tasks require higher-level thinking and/or problem-solving skills.
- ❑ Tasks often provide measures of metacognitive processes and attitudes as well as the more usual intellectual products.
- ❑ Tasks themselves represent meaningful instructional activities.
- ❑ Tasks represent real-world applications.
- ❑ Tasks are often time-consuming.
- ❑ Scoring rubrics are required.

Source: CSE/CRESST 1992

In a dramatic departure from the much-criticized traditional, single-answer, standardized testing, this new approach to assessment is intended to be so consistent with instructional goals that teaching *to* them would not only be acceptable, but would be considered exemplary practice.

Is Assessment the "Magic Band-Aid"?

Policy research does not give us cause for great optimism in using assessment to accomplish productive school improvement; over the last ten to twenty years, we have learned that it is very hard to use a top-down strategy to stimulate bottom-up reform. Where the changes really have to occur is in school, with teachers and students. Assessment research tells us we should be very careful in introducing assessment as a lever for change. What we find is that when there is a lot of pressure on test results and high-stake consequences follow from the results of tests, educational processes can get distorted. Traditional standardized testing has shown us important lessons that we need to address as we seek a better truth in alternative assessments:

▲ CAUTION ▲ *Narrowing of the curriculum can occur.*

John Cannell, a couple of years ago, reported that most states were scoring above the national average on nationally-normed standardized tests. If the concept of a norm-referenced test is that half the population scores above average, and half the population scores below average, it is a complete distortion to the meaning of those kinds of tests if everyone scores above average. It means the test is not measuring what it is supposed to measure and is not a very good indicator of student performance. Why does that happen? When CSE/CRESST looked more closely at it, we found that the pressure on educators to have their students do well on the test encouraged them to narrow their curriculum to teach only what was on the test. That was the best-case scenario of what was happening. In the worst case, teachers taught actual items from the test.

Under these conditions, what is the meaning of a test score? No one really cares whether a child can answer the specific fifty items that are on a one-hour test--that is really immaterial. What is important is that an hour's worth of measurement represents something larger. But if teachers are focusing only on those fifty items, the test ceases to represent something larger--it only represents those fifty items. So what do you have? You have the whole system focused on something fairly trivial.

▲ CAUTION ▲ *The system adjusts to protect children.*

Research done at CSE/CRESST on minimum competency testing also reveals that when there are high stakes associated with children's performance, the system adjusts to protect the children. For example, when children did not do very well on minimum competency tests and unacceptable numbers "failed," the standards were lowered so that almost all the children could pass. The good news is that the minimum competency notion did not hurt too many children; the bad news is that nothing really changed. It was largely, in many states, a symbolic effort. There was great hoopla around setting standards, developing tests, and having children perform well, but in most places 99.9 percent passed because the standards were set so that they *would* pass.

What is Happening Nationally?

Despite the lessons of assessment history and research, optimism in the power of assessment persists. Those who advocate testing most strongly look at the last ten years of research and say that assessment can motivate teachers to focus their instruction, but it focused their attention in the wrong direction--i.e., multiple-choice testing has forced teachers to focus on the lowest-level skills. These advocates assert that assessments need to encourage teachers to focus on the development of students' higher-level skills. In other words, the policy intent is right, but the test instrument is wrong; thus the move toward alternative assessment to improve instruction.

There are many groups and individuals advocating changes in assessment, and what they advocate includes varying levels of alternative assessment. It is a quite confusing national scene in which politics plays an active role and there is considerable overlap in who is actually doing what; for example, among the individuals with key roles in the various panels and groups listed below, Lauren Resnick and Mark Tucker have multiple involvements.

Major Players

America 2000

Context:	President Bush's agenda for education.
Mission:	Create American Achievement Tests to measure world class standards in five subjects (English, math, science, history, and geography).
Target:	Grades 4, 8, and 12.
Supports:	1. The use of these test results by colleges and employers for admission and hiring. 2. Standards and tests developed in conjunction with the National Education Goals Panel (NEGP).
Status:	Intend to initiate testing by 1993.

National Education Goals Panel (NEGP)

Context:	Formed by President and state governors in June 1990.
Mission:	Monitor progress on six national education goals.
Target:	Grades 4, 8 and 12.
Supports:	National standards and assessments in the five subjects.
Status:	Senate Bill 2 proposes that NEGP work to identify indicators to assess progress toward the national goals.

National Council on Education Standards and Testing (NCEST)

Context:	Created by governors, Congress, and the the Bush administration in June 1991.
Mission:	Examine the desirability and feasibility of developing national standards and a national assessment system.
Supports:	1. National standards with a national system of tests that is keyed (calibrated) to those standards. 2. State/local responsibility for test development and administration. 3. Delivery standards to assure students' equal opportunity to achieve.

Secretary's Commission on Achieving Necessary Skills (SCANS)

Context: A 31-member commission formed by former Labor Secretary Dole in 1990 and supported by current Labor Secretary Lynn Martin; is seen as an outcome of the Commission on the Skills of the American Workforce that was formed by the National Center on Education and the Economy in 1989.

Mission: Examine demands of workplace and determine whether students are prepared to meet them.

Target: Grades 8 and 12.

Supports: Cross-disciplinary competencies that students need for the world of work; based on a three-part foundation that includes thinking skills and personal qualities.

Status: It is expected that the federal government (through either the Dept. of Labor or Education) will soon issue a large RFP for test development.

National Standards Project

Context: Foundation-funded project lead by Lauren Resnick and Mark Tucker.

Mission: Develop "voluntary" national examination system.

Target: K-12 schools.

Supports:
1. Performance assessment, portfolios, and extended projects.
2. A single standard with equitable opportunity for all students.
3. A single national standard to which different tests would be calibrated (similar to NCEST).
4. Expansion, over time, to most subjects, including foreign language, fine arts, and workforce readiness.
5. Using the assessment by 2000 as the basis for college admissions, hiring, school incentives, and high school graduation.

Status:
1. 450 Teachers in 17 states are involved in pilot assessments in math, reading, and writing.
2. Ideas are strongly reflected in NEGP and NCEST reports.

National Assessment of Educational Progress (NAEP)

Context: Created by Congress in 1969, NAEP has been the only regularly conducted national survey of educational achievement at the elementary, middle, and high school levels, assessing student performance in reading, mathematics, science, writing, history/geography, and other fields.

Target: The Bush administration favors regular assessment in grades 4, 8, and 12 in the five subjects, and collection of state-by-state data.

Status:
1. Currently, NAEP information on group-level student performance is available to policymakers at the national, state, and local levels.
2. The Bush administration wants NAEP to enable reliable comparisons of states, districts, and schools within states; changes in congressional mandates would be required.

National Assessment Governing Board (NAGB)

Context: Mandated by Congress to provide policy direction to NAEP.

Mission: Set national standards for what students should know and be able to do as a basis for NAEP assessments.

Council of Chief State School Officers (CCSSO)

Context: Organization of states' chief education officers.
Mission: Work with others to coordinate a national assessment agenda.

Student Assessment Consortium (SAC)

Context: Initial collaboration between the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) and CCSSO.
Mission: Pool efforts across states to develop innovative assessments.
Status: Thirty states participating in work groups on math, social studies, reading, art, and workforce readiness.

Test Publishers

Commercial and other test publishers are developing a number of alternative assessment programs, some in conjunction with state testing programs. Some examples are:

- **The American College Testing Program (ACT)** is developing *Work Keys*, a national system for teaching and assessing employability skills that is directed at students who enter the workforce upon completing high school.
- **The College Board** is developing *Pacesetters*, which is modeled after the Advance Placement Test but intended for all 12th graders.
- **The Educational Testing Service (ETS)** is developing *Worklink*, an on-line record-keeping system designed to replace transcripts, and is considering developing a work force readiness assessment that can be keyed to *Worklink*.

How Much is This Going to Cost?

It is difficult to determine exactly what the cost of this new approach to assessment will be. Obviously, there will be sizable *development costs* involved to develop the new assessment measures being advocated. In contrast to traditional multiple-choice testing, new performance-based assessments will entail significant *ongoing administration and scoring costs*. CSE/CRESST conducted a set of interviews with state assessment directors about two years ago to see what states were doing in alternative assessment. Of course, the landscape has changed dramatically in the last two years, but at that point state assessment directors indicated that the regular multiple-choice tests cost them $1.50 per child. Their estimate for administering and scoring these new kinds of assessments was anywhere between $3.00 and $10.00 per child. Even the latter estimate is probably low and would vary according to the level of teacher involvement in the administration and scoring process.

Another argument being mounted is that we are not focusing only on assessment instruments when talking about new kinds of assessments, but also on changes in instruction. This re-thinking of assessment will require tremendous *staff development* in curriculum and instruction; this is where a lot of people are thinking the money is going to come from, in the hopes that at long last there will be some solid integration between assessment and instruction.

What are the Research and Development Challenges?

The alternative assessment "movement" has exciting potential, but having interesting and innovative tasks on a test instrument is no proof that the assessment is valid. We have to figure out how to validate these assessments and establish their reliability. This is a significant challenge, and it is important to be mindful of several points:

▲ CAUTION ▲ *There is a paucity of empirical research.*

CSE/CRESST is in the process of collecting data on what is going on in the states and local districts with the development process. In addition to examples and descriptions of what is being done in terms of subject matter, grade level and the like, CSE/CRESST asked people for any data related to the quality of the measures. In the great majority of cases there is no data on the quality of the measures. Practices are being called assessments, but much more data is needed to demonstrate that these are good assessments.

▲ CAUTION ▲ *We must re-think current psychometric theory.*

In alternative assessments, because each of the tasks is very time-consuming, there will not be a lot of samples of student work as in a multiple choice test. Consequently, it will be more difficult to talk in traditional terms regarding the reliability and validity of the measures. As a result, we need new measurement theory approaches. At the same time, it is not wise to ignore what is known about sound measurement as we go about defining new alternatives. Both assessment people and instructional people ought to be driving the action.

▲ CAUTION ▲ *Transferring and generalizing findings poses critical issues.*

One of the keys to alternative assessment is knowing what you have when you have the results of the assessment. People who have been working in this area have, for example, looked at children's performance on what is called "hands-on science experimentation." They examine the child's performance as well as how well the child's response can be generalized. If a child does one task well (e.g., designing an experiment about paper towels), what is the probability that he or she will do well on another hands-on science experiment? To get a reliable estimate of how well a child would do on the next task, you need a whole lot of samples on different kinds of hands-on experiences.

▲ CAUTION ▲ *Face validity is not sufficient.*

Just because a task looks like it measures higher-level thinking skills does not necessarily mean that it does. Just because a task looks interesting and innovative does not mean that it measures anything reliably or anything important. An assessment has to have more than face validity to qualify as a good assessment.

▲ CAUTION ▲ *We must stay attuned to the real consequences.*

Minimum competency tests and multiple-choice testing were not instituted because anyone wanted to harm schools, teachers, or students. They were instituted because policymakers thought they were going to be of help and benefit to the educational system. However, strange distortions occurred when there was tremendous pressure placed on test scores. Those same kinds of distortions can happen with these new kinds of measures as well. This is particularly important because student performance will be judged on the basis of fewer test items or tasks. On the one hand, assessment should not provide any surprises--everyone should know what is coming on the test, such as the skills tested and range of content covered. However, if students know exactly what is going to be on the test or if they are only provided instruction and practice in the exact things that are going to be on the test, a misuse of the system will result.

▲ CAUTION ▲ *Equity issues pose difficulties.*

Some stakeholders advocate alternative assessment as a means to provide more equity for students. In the short run, it is fairly clear that disadvantaged students are going to do worse on these kinds of measures than they do on traditional measures. If you follow the logic that says that standard multiple-choice testing measures have narrowed the curriculum to basic skills and have caused teachers to emphasize rote-level learning--worksheets, and the like--it is clear that the populations who have been exposed to this kind of teaching emphasis are not going to perform well initially on these new assessments. To a great extent, the children most exposed to this "narrowing" are currently in the schools where there is the most pressure to improve test scores (schools serving underachieving students). Those children, who over the last decade have had the least opportunity to develop higher-level thinking skills and the least opportunity to engage in these complex tasks, are not going to do well when presented with alternative assessments tasks. In fact, CSE/CRESST experience shows that when *any* child is presented with these tasks they do poorly.

▲ CAUTION ▲ *Calibration is tricky.*

At the federal level, we hear that there will be a national system of tests; everything will be calibrated to a single standard, but different tests can be developed at the local level. Yet we currently lack the technical capacity to accomplish such calibrations. An enormous problem will occur if we do not solve that issue, and we could find ourselves with a very inequitable system. If we have a national system of tests that we hope will be used for college admissions and hiring decisions, and if, for example, Montana develops an easy test while Georgia develops a difficult test, people in Georgia will have a much tougher time getting into school, obtaining jobs, and the like.

How Do We Ensure Quality Assessment?

Despite all the attention and enthusiasm about alternative assessment, most developers of these new assessments are at the design and prototype stages and are still some distance from tackling the critical issues of validity and quality assurance. The new alternative assessments being offered require an expanded set of criteria for determining their quality. Linn, Baker, and Dunbar (1991) provide the following helpful list of judgment criteria:

Quality Criteria for Alternative Assessment

Consequences	Does the tool have positive consequences, or are there unintended effects such as narrowing of curriculum, adverse effects on disadvantaged students, etc.?
Fairness	Does the tool consider fairly the cultural background of those students taking the test? Have all students had equal opportunity to learn the complex thinking and problem-solving skills which are the targets of these new assessments?
Transfer and Generalizability	Are the results reliable across raters and consistent in meaning across locales?
Cognitive Complexity	Does the tool actually assess higher-level thinking skills?
Content Quality	Does the tool reflect and draw on critical, enduring aspects of content?
Content Coverage	Do the assessment tasks represent a full curriculum?
Meaningfulness	Does the tool engage students in meaningful problems, resulting in worthwhile educational experiences and greater motivation for performance?
Cost and Efficiency	Has sufficient attention been given to efficient data-collection designs and scoring procedures?

Source: Linn, Baker and Dunbar (1991)

What Can We Expect to See in the Near Future?

There is considerable enthusiasm for some kind of national system of tests, with NAEP as the parallel assessment system. These tests could be funded by Congress, though the source of funding for actual development is unclear. Hopefully, money from the federal government will support some joint test-development efforts, potentially funneled through chief state school officers and their departments.

The hallmark of this national system of tests is that it will be keyed to national "world-class" standards in the subjects that are assessed. According to the federal vision, assessment instruments will differ among states, but somehow the states' different instruments will be calibrated to a single standard. Over the next two years, we are likely to see national standards developed in almost every content area. Additionally, we can expect to see considerable efforts to determine whether technical quality issues can be resolved.

Assessment has tremendous potential. However, at this point in time the rhetoric far surpasses the reality. We need to move forward intelligently, making sure that we don't expect assessment to do more than it can do.